A Method for Allocating Web Sites on a Web Server Cluster Based on Balancing Memory and Load Requirements

ABSTRACT

A method for operating a server cluster that includes N server nodes that service client requests. Each client request is directed to one of a plurality of sites hosted on the server cluster. Each site is identified by a domain name, and each server node is identified by an address on a network connecting the clients to the server nodes. The computational resources required to service the requests to each of the sites over a first time period are measured and used to group the sites into N groups. Each group is assigned to a corresponding one of the server nodes. The groups are chosen such that, for each pair of groups, the difference in the sum of the measured computational resources is within a first predetermined error value. Configuration information defining a correspondence between each of the sites and one or more of the server nodes assigned to the groups containing that site is then provided to a router accessible from the network. The groupings are periodically updated by measuring the computational resources required to service the requests to each of the sites over a second time period; and grouping the sites into N new groups. The new groups are constructed by swapping sites between the previous groups. The new groups are constructed such that, for each pair of new groups, the difference in the sum of the measured computational resources over the second time period is within a second predetermined error value. The new grouping that satisfies the second error condition and for which the new groups differ from the previous groups by as few site swaps as possible is preferred.